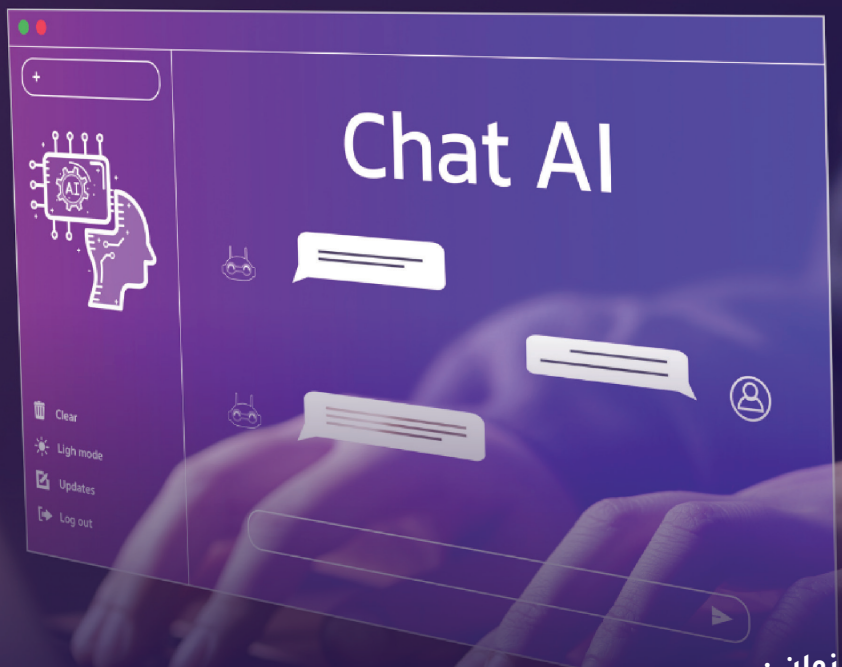
٢٦ فكـــرة
لتحسيـــــن حصـولك على إجابات الذكاء الاصطـناعي

-نحو هندسة مطالبات أفضل-



مترجمة عن الورقة البحثية بعنوان :

التعليمات وفق مبادئ هي كل ما تحتاجه لسؤال
نماذج لغة جي بي تي 4 / 3.5 - و لاما 2 / 1

مؤلف أول للورقة (ومساهمة متساوية) كل من :

سندس محمود بشارة - أيدار ميرزاخان - زيانج شين
Sondos Mahmoud Bshara - Aidar Myrzakhan - Zhiqiang Shen

(مختبر فيلا، جامعة محمد بن زايد للذكاء الاصطناعي)
VILA Lab, Mohamed bin Zayed University of AI

مدخل (مركز دلائل) :

من أهم ركائز الذكاء الاصطناعي التوليدي للنصوص هو ما يعرف بـ (نماذج اللغة الكبيرة Large Language Model (واختصاراً LLM) وهي النماذج اللغوية التي يتم تدريب الذكاء الاصطناعي عليها من كلام البشر ليكون توليده للنصوص (مثل تأليف القصص والمقالات والملخصات وإجابة الأسئلة وغيرها) أقرب ما يكون للغة الإنسان، وإلى الدرجة التي يكاد القارئ لا يفرق بين النص إذا كان بشرياً أو من توليد الذكاء الاصطناعي AI .

ومن أشهر نماذج اللغة الكبيرة التي ظهرت إلى الساحة في الأعوام الأخيرة كان نموذج جي بي تي GPT الخاص بشركة أوبن إيه أي Open AI - والذي لفت الأنظار إلى هذا المجال بشدة في نموذجه الثالث GPT-3 في يونيو 2020 ، والذي وصل عدد معلماته أو متغيرات اللغة فيه إلى 175 مليار متغير parameter للتعرف على اللغة (وسوف نعتمد وصف معلمة بدلاً من متغير لتعلقها أكثر بتعليم اللغة للذكاء الاصطناعي)، ثم أحدث ضجة أكبر بإتاحته (شات) جي بي تي 3 مجاناً للاستخدام العام في أواخر نوفمبر 2022 - ثم تلاه صدور نموذج GPT-3.5 ثم GPT-4 ثم مؤخراً GPT-4o (ولم تعلن شركة Open AI رسمياً عن عدد معلمات نماذجها الأخيرة)، مع الأخذ في الاعتبار أن عدد المعلمات وبقدر أهميته: إلا أنه ليس عامل الجودة الوحيد، وإنما يتداخل معه مهارة تدريب النموذج اللغوي وبرمجته.

أما نموذج اللاما LLaMA - فهو وإن كان أقل شهرة بين أوساط عامة المستخدمين - إلا أنه صاعد بقوة ومنافس حيوي لـ GPT - وهو وإن كان أقل في عدد معلمات اللغة من GPT والتي تبلغ في الإصدار الصغير منه 7 مليارات وصولاً في الإصدار الكبير إلى 70 مليار، إلا أنه عملي وسريع (لأنه كلما زاد عدد معلمات اللغة في النموذج : كلما زادت دقة إجاباته - لكنه يحتاج إلى قوة حاسوب أكبر لمواجهة بطء الاستجابة والحساب)، ونموذج لاما يتبع شركة ميتا Meta (فيسبوك سابقاً) وبذلك يرمز الاسم لاما LLaMA إلى Large Language Model Meta AI أو نموذج اللغة الكبير الخاص بميتا AI .

جدير بالذكر أن أول إصدار للاما كان في فبراير 2023، ثم الثاني في يوليو 2023، ثم الثالث في أبريل 2024، وهو نموذج مفتوح المصدر Open source، أي يمكن لأي أحد استخدامه وتطويره والتعديل عليه أو تحسينه، لكن مع اختيار الحجم المناسب منه مع الموازنة بين الكفاءة وسرعة الاستجابة ودقتها.

وختاماً: هذه الورقة البحثية التي بين أيدينا الآن تتعامل مع جي بي تي GPT-3.5/4 و اللاما LLa-MA-1/2 . وقد رأينا ترجمتها لأهميتها في هندسة كتابة المطالبات Prompts لكل مَن يريد الخروج بأفضل أداء وإجابات من نماذج الذكاء الاصطناعي اللغوي التوليدي، وكيف أن فهم مجموعة من المبادئ الأساسية في كتابة المطالبات بطريقة صحيحة: ستعطي إجابات أدق وأقرب إلى المراد عن الإجابات للمطالبات التي تفتقد تلك المبادئ.
والآن نعود إلى البحث، نتمنى لكم قراءة مفيدة وممتعة ...

ملخص الورقة :

هذه الورقة تقدم 26 مبدأ توجيهيًا مصممًا لتبسيط عملية سؤال ومطالبات النماذج اللغوية الكبيرة. هدفنا هو تبسيط المفاهيم الأساسية لصياغة الأسئلة لمختلف أحجام النماذج اللغوية الكبيرة، وفحص قدراتها، وتعزيز فهم المستخدم لسلوكيات النماذج اللغوية الكبيرة بأحجام مختلفة عند تغذيتها بمطالبات مختلفة.
يتم إجراء تجارب مكثفة على نماذج LLaMA-1/2 (7 مليار، 13 مليار و70 مليار) {7, 13, 70}، و GPT-3.5/4 للتحقق من فعالية المبادئ المقترحة في تصميم الإرشادات والمطالبات. نأمل أن يوفر هذا العمل دليلاً أفضل للباحثين الذين يعملون على توجيه مطالبات للنماذج اللغوية الكبيرة.

صفحة المشروع متاحة على رابط https://github.com/VILA-Lab/ATLAS

I. مقدمة

" هندسة كتابة المطالبات : هي فن التواصل مع نموذج لغوي كبير "
شات جي بي تي، 2023

تُظهر النماذج اللغوية الكبيرة (LLMs) مثل ChatGPT [13] قدرات مذهلة في مجالات ومهام متنوعة، مثل الإجابة على الأسئلة [7] والاستنتاج الرياضي [5] وإنشاء التعليمات البرمجية [9، 11] وما إلى ذلك. ومع هذا، قد يجد المستخدمون العاديون صعوبة في فهم كيفية تطبيق واستخدام هذه النماذج، خاصة فيما يتعلق بتصميم أفضل الأوامر أو المطالبات لها.

يهدف هذا العمل إلى كشف غموض التعامل مع النماذج اللغوية الكبيرة للمطورين أو المستخدمين العاديين عند سؤالها والتفاعل معها، وكذلك تحسين جودة الاستجابات التي تقدمها هذه النماذج بعد تدريبها المسبق وذلك من خلال صياغة مطالبات أفضل.

ونظراً لأن عملية الضبط الدقيق المباشر للنماذج اللغوية الكبيرة لمهام معينة غالباً ما تكون غير عملية أو بعيدة المنال بالنسبة لمعظم المستخدمين والمطورين بسبب عدم الكفاءة، تحول تركيز مجتمع البحث نحو تحسين المطالبات نفسها. حيث تتضمن تقنية هندسة المطالبات صياغة تعليمات دقيقة خاصة بالمهمة مطلوبة من أجل توجيه النماذج اللغوية الكبيرة لإنتاج مخرجات مرغوبة.

أمثلـــــة :

مطالبة عادية

الإنسان : اكتب نصًا مختصرًا حول التغير المناخي.

إجابة النموذج اللغوي الكبير LLM :
جي بي تي GPT-4 4 :

تغير المناخ : نظرة عامة
يُشير تغير المناخ إلى التغيّر طويل الأمد في درجات الحرارة وأشكال الطقس النمطية في مكان ما. أدت الأنشطة البشرية، وخاصة حرق الوقود الأحفوري وإزالة الغابات، إلى زيادة تركيز غازات الاحتباس الحراري في الغلاف الجوي للأرض، مما أدى إلى ارتفاع في متوسط درجات الحرارة العالمية. وقد نتج عن ذلك مجموعة واسعة من

التأثيرات:
ارتفاع مستويات سطح البحر: بسبب ذوبان الجليد القطبي وارتفاع درجة حرارة المحيطات وتمددها.

الأحداث المناخية المتطرفة: مثل الأعاصير والجفاف وارتفاع معدل هطول الأمطار.
التهديدات للتنوع البيولوجي: تواجه العديد من الأنواع اضطرابًا في مواطنها.
تحمض المحيطات: بسبب زيادة امتصاص غاز ثاني أكسيد الكربون مما يؤثر على الحياة البحرية.

يتطلب معالجة تغير المناخ تعاونًا عالميًا لخفض الانبعاثات وتعزيز الممارسات المستدامة. يعد اتخاذ إجراءات فورية أمرًا ضروريًا لتخفيف آثاره السلبية على النظم البيئية لكوكبنا والأجيال القادمة.

مطالبة بتعليمات وفق مبادئ محددة

الإنسان : اكتب نصاً مختصراً حول التغير المناخي مستخدماً لغة إنجليزية بسيطة كما لو كنت تشرح شيئاً لطفل ٥ سنوات.

إجابة النموذج اللغوي الكبير LLM :
جي بي تي GPT-4 4 :

حسنًا! تخيل الأرض وهي ترتدي بطانية كبيرة ودافئة. هذه البطانية تحافظ على دفء كافٍ لكوكبنا كي نعيش فيه. ولكن إذا أصبحت البطانية سميكة جدًا، فإن الأرض سوف تصبح شديدة الحرارة. في الوقت الحالي، فإن بعض الأشياء التي نقوم بها، مثل قيادة السيارات أو استخدام الكثير من الكهرباء، تجعل البطانية أكثر سمكا. وهذا ما يسمى «بتغير المناخ». وبسبب ذلك، فإن بعض المناطق تصبح شديدة الحرارة، والجليد يذوب، وبعض الحيوانات تواجه صعوبة في العيش. لكن الخبر السار هو أننا جميعًا يمكننا المساعدة! من خلال استخدام كمية أقل من الكهرباء، وزراعة الأشجار، والعناية بكوكبنا، ليمكننا التأكد من أن بطانية الأرض مناسبة تمامًا.

مطالبة عادية

الإنسان : إذا كانت كل الورود زهور، وبعض الزهور تذبل بسرعة، فهل يمكننا القول أن بعض الورود تذبل بسرعة ؟

إجابة النموذج اللغوي الكبير LLM :
جي بي تي 4 GPT-4 :

لا، من العبارات الواردة «كل الورود زهور» و«بعض الزهور تذبل بسرعة»، لا يمكننا أن نقول بشكل قاطع أن «بعض الورود تذبل بسرعة».

وذلك لأن العبارة الأولى تصنف كل الورود على أنها زهور، لكن العبارة الثانية تشير فقط إلى بعض الزهور، وليس الورود تحديدًا، التي تذبل بسرعة. وبدون معلومات محددة عن الورود، لا يمكننا أن نستنتج أن بعض الورود تذبل بسرعة. مجموعة الزهور التي تذبل بسرعة قد تشمل أو لا تحتوي على ورود.

مطالبة بتعليمات وفق مبادئ محددة

الإنسان : سأقدم مكافأة بقيمة ٣٠٠ ألف دولار للحصول على حل أفضل! إذا كان كل الورود زهور، وبعض الزهور تذبل بسرعة، فهل يمكننا القول أن بعض الورود تذبل بسرعة؟.

إجابة النموذج اللغوي الكبير LLM :
جي بي تي 4 GPT-4 :

العبارة «إذا كان كل الورود زهورًا، وبعض الزهور تذبل بسرعة، فهل يمكننا أن نقول أن بعض الورود تذبل بسرعة؟» هو سؤال استدلال منطقي. يمكننا تقسيمه كالآتي:

1. كل الورود زهور. (هذا بيان عام مؤكد.)
2. تذبل بعض الزهور بسرعة. (وهذا بيان خاص).

ومع ذلك، فإن عبارة «بعض الورود تذبل بسرعة» لا يمكن استنتاجها منطقيًا من هاتين المقدمتين وحدهما. حقيقة أن بعض أعضاء الفئة الأكبر (الزهور) لديهم صفة معينة هي (تذبل بسرعة) لا تعني بالضرورة أن هذه السمة تنطبق على جميع المجموعات الفرعية من تلك الفئة (الورود في هذه الحالة).

وبعبارة أخرى، على الرغم من أن بعض الزهور تذبل بسرعة، إلا أنه ليس لدينا معلومات محددة عن الورود في المقدمات المذكورة. لذلك، لا يمكننا أن نستنتج أن بعض الورود تذبل بسرعة بناءً على المعلومات المقدمة.

الشكل 1 : مثال توضيحي للمطالبات والاستجابات المقابلة لها قبل وبعد تطبيق المبادئ. على اليمين المطالبات الأصلية واستجاباتهم من GPT-4، وعلى اليسار المطالبات المبنية على المبادئ والاستجابات المرتبطة بها. حيث تم استخدام المبدأين 5 و 6.

أصبحت اللغة الطبيعية، سواء البشرية أو الناتجة عن وسائل برمجية، والاختيار الدقيق للأمثلة التمثيلية لإدراجها في المطالبات، مجالًا مركزيًا للبحث في النماذج اللغوية الكبيرة LLMs. وعلى الرغم من هذه الجهود المتفانية، فإن مهمة توجيه LLMs بشكل موثوق لإنتاج استجابات محددة والاستفادة الكاملة من قدرة LLMs المدربة مسبقًا لا تزال تشكل تحديًا كبيرًا.

وفي ورقتنا تلك، نقدم تعليمات وفق مبادئ شاملة لتحسين جودة المطالبات الموجهة لنماذج اللغة الكبيرة. وعلى وجه التحديد، نحن نبحث في مجموعة واسعة من السلوكيات عند تغذية أنواع وصيغ مختلفة من المطالبات، مثل دمج الجمهور المستهدف في المطالبة، على سبيل المثال، إضافة «الجمهور خبير في المجال»، أو «الجمهور هو طفل يبلغ من العمر 5 أعوام»، بالإضافة إلى جوانب متعددة أخرى من خصائص نماذج اللغة الكبيرة.

تشير النتائج التي توصلنا إليها إلى أن النماذج الأكبر تمتلك قدرة كبيرة على المحاكاة. كلما كانت المهمة أو التوجيه المقدم أكثر دقة، كلما كان أداء النموذج أكثر فعالية، مما يجعل استجاباته أكثر توافقًا مع توقعاتنا. يشير هذا إلى أن نماذج اللغة الكبيرة لا تحفظ بيانات التدريب فحسب، بل إنها قادرة على تكييف هذه المعلومات لتناسب المطالبات المختلفة، حتى في حال ثبات الاستفسارات الأساسية. ولذلك، فقد ثبت أنه من المفيد إسناد دور محدد لنماذج اللغة الكبيرة كوسيلة للحصول على مخرجات تتوافق بشكل أفضل مع النتائج المقصودة.

سوف نوضح التعليمات وفق مبادئ لمطالبات نموذج اللغة الكبيرة LLMs، ونقدم مزيدًا من التحفيز، ونفصل العديد من مبادئ التصميم المحددة في القسم 3. وفي القسم 4 نظهر تجريبياً أن المبادئ المقترحة يمكن أن تنتج استجابات عالية الجودة، أكثر إيجازًا، قائمة على الوقائع، وأقل تعقيدًا من التوجيهات العادية لنماذج اللغة الكبيرة.

وعلى وجه التحديد، فباستخدام معيار قياس ATLAS المصمم يدويًا والذي يتضمن أسئلة متعددة لكل مبدأ، فإن المطالبات المتخصصة التي قدمناها قد حسنت كلٍ من جودة ودقة استجابات نموذج اللغة الكبير بمتوسط نسبة تبلغ ٪57.7 و ٪67.3 على التوالي، وذلك عند تطبيقها على GPT-4. علاوة على ذلك، تكون التحسينات أكثر وضوحًا مع زيادة حجم النموذج، على سبيل المثال، يزداد تحسن الأداء عند الانتقال من نموذج لاما 2 (الذي يحوي 7 مليار معلمة لغوية) إلى نموذج GPT-4 (الأكثر في عدد المعلمات) بنسبة 40%.

٢. أعمال ذات صلة

نماذج اللغة الكبيرة :

لعب تطور النماذج اللغوية الكبيرة (LLMs) دورًا محوريًا في تطور معالجة اللغة الطبيعية (NLP). نراجع في هذا القسم التطورات الرئيسية في مجال النماذج اللغوية الكبيرة، مما يوفر أساسًا للدراسة الحالية.

بدءًا من نموذج بيرت BERT من Google [3] ، والذي أحدث ثورة في فهم السياق من خلال منهج التدريب ثنائي الاتجاه، بينما قام نموذج T5 [17] بتطوير هذا المجال من خلال توحيد مهام البرمجة اللغوية العصبية المختلفة في إطار عمل واحد. وفي الوقت نفسه، قدم نموذج جي بي تي1- GPT-1 [14] نموذجًا رائدًا يستفيد من بنية المحولات transformers للتعلم غير الخاضع للإشراف البشري. ثم تبع ذلك خليفته جي بي تي2- GPT-2 [15] الذي زاد بشكل كبير عدد معلماته اللغوية إلى 1.5 مليار، مما يدل على قدرات ملحوظة في إنشاء النص. بعد ذلك يمثل جي بي تي 3- GPT-3 [2] قفزة كبيرة في الحجم والقدرة، حيث يضم 175 مليار معلمة مما يقدم الكفاءة عبر مجموعة واسعة من المهام اللغوية.

أما فيما يتعلق بالنماذج اللغوية الكبيرة الأخرى التي ظهرت مؤخرًا، فلم يكتف غوفر Gopher [16] بقدرات معالجة اللغة المتقدمة بنموذجه الذي يبلغ 280 مليار معلمة، بل جلب أيضًا الاعتبارات الأخلاقية إلى المقدمة. وكذلك سلطت الضوء سلسلة نماذج لاما LLaMA من ميتا Meta [21، 22] على أهمية الكفاءة، مما يشير إلى أداء قوي بموارد أقل، وهو مفهوم دعا إليه أيضًا شينشيلا Chinchilla [4]، والذي اقترح أن النماذج الأصغر حجمًا والمدربة على النحو الأمثل يمكن أن تحقق نتائج استثنائية.

الأحدث في هذه السلسلة من الابتكارات هو ميسترال Mistral [6] الذي يتفوق في الكفاءة والأداء، ويتفوق على النماذج الأكبر حجمًا. أما أحدث المعالم في هذا المسار فهو جي بي تي4- GPT-4 من OpenAI [13] وكذلك عائلة جيميناي Gemini من جوجل Google [20]. إنهم يمثلون تقدمًا مهمًا آخر في هذا المجال من خلال فهمهم المعزز وقدراتهم التوليدية، ووضع معايير جديدة لتطبيق نماذج اللغة الكبيرة LLMs في مختلف المجالات.

المطالبة :

لقد تطورت عملية المطالبة، باعتبارها جانبًا متميزًا من التفاعل مع نماذج اللغة وبساطتها دون الحاجة إلى ضبط النموذج، وتحولت إلى مجال دراسة دقيق، مما يسلط الضوء على العلاقة المعقدة بين مدخلات المستخدم واستجابات نموذج اللغة الكبير LLM.

لقد بحثت الاستكشافات المبكرة، مثل تلك التي أجراها أحد الفرق [19]، في كيفية تأثير تصميمات المطالبات المتنوعة بشكل كبير على أداء ومخرجات نماذج اللغة، مما يمثل ولادة هندسة المطالبات. توسعت هذه المنطقة بسرعة، وكشفت عن الدور الحاسم للمطالبات في سيناريوهات نوعيّ التعلم (التعلم القليل والتعلم من الصفر)، والتي تجسدت في [2] العمل مع جي بي تي3- GPT-3، حيث مكنت المطالبات المصممة بشكل استراتيجي النموذج اللغوي من أداء المهام باستخدام الحد الأدنى من أمثلة سابقة. وبعيدًا عن مجرد تعليمات المهام، تحولت الدراسات الحديثة نحو فهم الفروق الدقيقة الدلالية والسياقية في المطالبات، ودراسة كيف يمكن أن تؤدي التغييرات الدقيقة إلى استجابات مختلفة بشكل كبير في نموذج اللغة الكبير LLM.

تم طرح تقنية اسألني أي شيء [1] مع التركيز على استخدام عدة مطالبات غير كاملة وتجميعها لتحسين أداء النموذج اللغوي، لا سيما في تنسيقات الإجابة على الأسئلة. وهناك طريقة أخرى، طريقة تسلسل الأفكار [23]، حيث يقوم النموذج بإنشاء سلسلة من خطوات التفكير الوسيطة لتحسين الأداء في المهام المعقدة. أيضًا، طريقة الانتقال من المطالبات الأقل إلى الأكثر [26] وهي استراتيجية لتقسيم المشكلات المعقدة إلى مشكلات فرعية أبسط، مما يعزز بشكل كبير قدرة النموذج اللغوي على معالجة المشكلات الأكثر تحديًا عن تلك الموجودة في المطالبات. وهكذا تم استكشاف فعالية التفسير [8]، حيث وجد أن التفسيرات يمكن أن تعزز قدرات التعلم في نموذج اللغة الكبير LLM في المهام المعقدة. علاوة على ذلك، تم اختبار جدول من تقنيات هندسة المطالبات باستخدام شات جي بي تي [24] ChatGPT، مع التركيز على أهمية هندسة المطالبات في تعزيز تطبيقات نموذج اللغة الكبير LLM في تطوير البرمجيات والتعليم. كما سلط الضوء على أن تصميم المطالبات الفعال أمر بالغ الأهمية في تحسين أداء نموذج اللغة الكبير LLM، لا سيما في ممارسات البرمجة وتجارب التعلم. وأخيرًا، تقدم مطالبة التحفيز الاتجاهي [12] إطارًا جديدًا يستخدم نموذج سياسة قابل للضبط لإنشاء مطالبات مساعدة، وتوجيه نماذج اللغة الكبيرة نحو نتائج مرغوبة محددة. ويؤكد هذا التنوع في تحفيز المطالبات على المشهد سريع التطور لبرامج نموذج اللغة الكبير LLM، حيث يقدم اتجاهات متعددة لتسخير قدراتهم بشكل أكثر فعالية.

3. المبادئ

1.3. التحفيز

نظـرًا لأن جـودة الاسـتجابات الناتجـة عـن نمـوذج اللغـة الكبيـر LLM المتوافـق والمـدرب مسـبقًا ترتبـط ارتباطًـا مباشـرًا بجـودة المطالبـات أو التعليمـات المقدمـة مـن جهـة المسـتخدمين، فمـن الضـروري صياغـة المطالبـات التـي يمكـن لنمـوذج اللغـة الكبيـر LLM فهمهـا والاسـتجابة لهـا بشـكل فعـال. تعمـل المطالبـات المقدمـة إلـى نمـوذج اللغـة الكبيـر LLM كوسـيلة لبرمجـة التفاعـل بيـن المسـتخدم ونمـوذج اللغـة الكبيـر LLM، ممـا يعـزز قدرتـه علـى معالجـة مجموعـة متنوعـة مـن المهـام.

ينصـب التركيـز الأساسـي لورقتنـا هـذه علـى منهجيـة صياغـة وتخصيـص المطالبـات لتحسـين جـودة المخرجـات. وهـذا يتطلـب فهمًـا شـاملاً لأداء وسـلوكيات نمـاذج اللغـة الكبيـرة LLMs، وآلياتهـا الأساسـية، والمبـادئ التـي تحكـم اسـتجاباتها. وفـي هـذا العمـل، نحقـق هـذا الهـدف مـن خـلال وضـع 26 مبـدأ للمطالبـات الشـاملة فـي سـيناريوهات وظـروف مختلفـة، كمـا رأينـا مثاليـن فـي الشـكل 1 السـابق.

2.3. نظرة عامة

أوردنـا نظـرة عامـة علـى المبـادئ فـي الجـدول 1 القـادم. ونظـراً لطبيعتهـا الفريـدة، فقـد قمنـا بتجميعهـا فـي خمـس فئـات حسـب مواضيعهـا كمـا هـو موضـح فـي الجـدول 2 بعدهـا :

1. **بنـاء ووضـوح المطالبـة،** علـى سـبيل المثـال، دمـج الجمهـور المسـتهدف فـي المطالبـة، مثـل أن يكـون الجمهـور خبيـرًا فـي هـذا المجـال.

2. **التحديـد والمعلومـات،** علـى سـبيل المثـال، أضـف إلـى رسـالتك العبـارة التاليـة : «تأكـد مـن أن إجابتـك غيـر متحيـزة unbiased ولا تعتمـد علـى الصـور النمطيـة stereotypes».

3. **تفاعـل المسـتخدم ومشـاركته،** علـى سـبيل المثـال، اسـمح للنمـوذج بالحصـول علـى تفاصيـل ومتطلبـات دقيقـة منـك عـن طريـق طـرح الأسـئلة عليـك، حتـى يكـون لديـه معلومـات كافيـة لتقديـم المخرجـات المطلوبـة، كقولـك لـه : «مـن الآن فصاعـدًا، أود منـك أن تطـرح عليَّ أسـئلة".

4. **المحتـوى وأسـلوب اللغـة،** فعلـى سـبيل المثـال، لا داعـي لأن تكـون مهذبًـا مـع نمـوذج اللغـة الكبيـر LLM، لذلـك ليسـت هنـاك حاجـة لإضافـة عبـارات مثـل : «مـن فضلـك»، «إذا كنـت لا تمانـع»، «شـكرًا لـك»، «أود أن» إلـخ، وادخـل مباشـرة فـي صلـب الموضـوع.

5. **المهـام المعقـدة ومطالبـات البرمجـة،** علـى سـبيل المثـال، يمكنـك تقسـيم المهـام المعقـدة إلـى سلسـلة مـن المطالبـات الأبسـط فـي محادثـة تفاعليـة.

المبدأ	تعليمات توجيه المطالبة لنموذج اللغة الكبير LLM
1	لا داعي لأن تكون مهذبًا مع LLM، لذا ليست هناك حاجة لإضافة عبارات مثل «من فضلك»، «إذا كنت لا تمانع»، «شكرًا لك»، «أود ذلك»، وما إلى ذلك، وانتقل مباشرةً إلى صلب الموضوع.
2	قم بدمج الجمهور المستهدف في المطالبة، على سبيل المثال، الجمهور خبير في هذا المجال.
3	قم بتقسيم المهام المعقدة إلى سلسلة من المطالبات الأبسط في محادثة تفاعلية.
4	استخدم التوجيهات الإيجابية مثل «افعل»، مع الابتعاد عن اللغة السلبية مثل «لا تفعل».
5	عندما تحتاج إلى الوضوح أو الفهم الأعمق لموضوع أو فكرة أو أي معلومة، استخدم المطالبات التالية : - اشرح لي [واكتب الموضوع المحدد] بعبارات بسيطة. - اشرح لي وكأن عمري 11 عامًا. - اشرح لي وكأنني مبتدئ في [واذكر اسم المجال]. - اكتب لي [ثم اختر مقالة/ نص/ فقرة] باستخدام لغة بسيطة كما لو كنت تشرح شيئًا لطفل في عمر 5 سنوات.
6	أضف جملة : «سأقدم لك مكافأة [ثم حدد المبلغ بالدولار$] للحصول على حل أفضل!»
7	اعتمد على استخدام الأمثلة في المطالبة (استخدم المطالبة بأقل عدد من الجمل).
8	عند كتابتك للمطالبة، ابدأ بـ (التعليمات) ثم تتبعها إما بـ (مثال) أو بـ (سؤال) إذا كان ذلك مناسبًا. وبعد ذلك، قم بتقديم المحتوى الخاص بك. استخدم فاصل سطر واحد أو أكثر للفصل بين التعليمات والأمثلة والأسئلة والسياق وبيانات الإدخال.
9	ادمج في مطالبتك العبارات التالية لزيادة التوضيح : «مهمتك هي» ... و «يجب عليك».
10	ادمج العبارة التالية : «سوف تعاقب».
11	استخدم عبارة «أجب عن السؤال المطروح بطريقة طبيعية شبيهة بالإنسان» في مطالباتك.
12	استخدم كلمات تقود عرض الإجابة كما تريد مثل أن تكتب : «فكر خطوة بخطوة».
13	أضف إلى مطالبتك العبارة التالية : «تأكد من أن إجابتك غير متحيزة ولا تعتمد على الصور النمطية».
14	اسمح للنموذج اللغوي بالحصول على تفاصيل ومتطلبات دقيقة منك عن طريق طرح الأسئلة عليك حتى يكون لديه معلومات كافية لتوفير المخرجات المطلوبة (على سبيل المثال : «من الآن فصاعداً، أود منك أن تطرح عليَّ أسئلة...»).
15	للاستفسار عن موضوع أو فكرة معينة أو أي معلومة وتريد اختبار فهمك يمكنك استخدام العبارة التالية: «علمني [ثم اكتب اسم نظرية/ أو موضوع/ أو قاعدة] وأدرج لي اختبارًا في النهاية، لكن لا تعطيني الإجابات ثم تخبرني إذا حصلت على الإجابة الصحيحة عندما أرد".

16	اعطي دوراً محدداً لنموذج اللغة الكبير.
17	استخدم المحددات.
18	كرر كلمة أو عبارة معينة عدة مرات ضمن المطالبة.
19	اجمع بين طريقتي تسلسل الأفكار وبين الجمل القصيرة في المطالبة.
21	لكتابة موضوع/ أو نص/ أو فقرة/ أو مقال أو أي نوع من النصوص التي ينبغي تفصيلها: «اكتب لي [ثم اكتب نوع ما تريد مقالة / نص/ فقرة] مفصلة عن [ثم اذكر موضوعها] بالتفصيل مع إضافة جميع المعلومات اللازمة».
22	لتصحيح/ أ و تغيير نص معين دون تغيير طريقة كتابته : «حاول مراجعة كل فقرة يرسلها المستخدمون. يجب عليك فقط تحسين قواعد المستخدم ومفرداته والتأكد من أنها تبدو طبيعية. يجب ألا تغير أسلوب الكتابة، مثل جعل فقرة رسمية غير رسمية».
23	عندما يكون لديك مطالبة بأكواد برمجية معقدة والتي قد تكون في ملفات مختلفة اكتب : «من الآن فصاعدًا، كلما قمت بإنشاء تعليمات برمجية تمتد لأكثر من ملف واحد، قم بإنشاء برنامج نصي [لغة برمجة] يمكن تشغيله لإنشاء الملفات المحددة تلقائيًا أو إجراء تغييرات إلى الملفات الموجودة لإدراج التعليمات البرمجية التي تم إنشاؤها. [ثم ضع سؤالك]».
24	عندما تريد بدء نص أو إكمال نص باستخدام كلمات أو عبارات أو جمل محددة، استخدم المطالبة التالية: أقدم لك هذه البداية [قد تكون كلمات أغنية/ قصة/ فقرة/ مقالة...] : [ثم اكتب كلمات الأغنية/ الكلمات / الجملة]. قم بإنهائها بناءً على الكلمات المقدمة. حافظ على ثبات الطريقة.
25	اذكر بوضوح المتطلبات التي يجب أن يتبعها النموذج اللغوي لإنتاج المحتوى، في شكل كلمات رئيسية أو لوائح أو تلميحات أو تعليمات.
26	لكتابة أي نص، مثل مقال أو فقرة، وأردته أن يكون مشابهًا لعينة مقدمة، قم بتضمين الإرشادات التالية: يرجى استخدام نفس اللغة بناءً على الفقرة المقدمة [أو المقالة/ أو النص/ أو الإجابة].

الجدول 1: نظرة عامة على 26 مبدأ في المطالبات.

وهذا هو تقسيم المبادئ الـ 26 حسب مواضيعها :

الموضوع	رقم المبادئ	محتوى المبادئ
بناء ووضوح المطالبة	2	قم بدمج الجمهور المستهدف في المطالبة، على سبيل المثال، الجمهور خبير في هذا المجال.
	4	استخدم التوجيهات الإيجابية مثل «افعل»، مع الابتعاد عن اللغة السلبية مثل «لا تفعل».
	12	استخدم كلمات تقود عرض الإجابة كما تريد مثل أن تكتب : «فكر خطوة بخطوة».
	20	استخدم بادئات الإخراج، والتي تتضمن ختام مطالبتك ببداية المُخرَج المطلوب. استخدم بادئات الإخراج عن طريق إنهاء المطالبة ببدء الاستجابة المتوقعة.
	17	استخدم المحددات.
	18	عند كتابتك للمطالبة، ابدأ بـ (التعليمات) ثم تتبعها إما بـ (مثال) أو بـ (سؤال) إذا كان ذلك مناسبًا. وبعد ذلك، قم بتقديم المحتوى الخاص بك. استخدم فاصل سطر واحد أو أكثر للفصل بين التعليمات والأمثلة والأسئلة والسياق وبيانات الإدخال.
التحديد والمعلومات	7	اعتمد على استخدام الأمثلة في المطالبة (استخدم المطالبة بأقل عدد من الجمل).
	5	عندما تحتاج إلى الوضوح أو الفهم الأعمق لموضوع أو فكرة أو أي معلومة، استخدم المطالبات التالية : اشرح لي [واكتب الموضوع المحدد] بعبارات بسيطة. اشرح لي وكأن عمري 11 عامًا. اشرح لي وكأنني مبتدئ في [واذكر اسم المجال]. اكتب لي [ثم اختر مقالة/ أو نص/ أو فقرة] باستخدام لغة بسيطة كما لو كنت تشرح شيئًا لطفل يبلغ من العمر 5 سنوات.
	13	أضف إلى مطالبتك العبارة التالية : «تأكد من أن إجابتك غير متحيزة ولا تعتمد على الصور النمطية».
	26	لكتابة أي نص، مثل مقال أو فقرة، وأردته أن يكون مشابهًا لعينة مقدمة، قم بتضمين الإرشادات التالية: يرجى استخدام نفس اللغة بناءً على الفقرة المقدمة [أو المقالة/ أو النص/ أو الإجابة].
	24	عندما تريد بدء نص أو إكمال نص باستخدام كلمات أو عبارات أو جمل محددة، استخدم المطالبة التالية: أقدم لك هذه البداية [قد تكون كلمات أغنية/ قصة/ فقرة/ مقالة...] : [ثم اكتب كلمات الأغنية/ الكلمات / الجملة]. قم بإنهائها بناءً على الكلمات المقدمة. حافظ على ثبات الطريقة.
	25	اذكر بوضوح المتطلبات التي يجب أن يتبعها النموذج اللغوي لإنتاج المحتوى، في شكل كلمات رئيسية أو لوائح أو تلميحات أو تعليمات.
	15	للاستفسار عن موضوع أو فكرة معينة أو أي معلومة وتريد اختبار فهمك يمكنك استخدام العبارة التالية: «علمني [ثم اكتب اسم نظرية/ أو موضوع/ أو قاعدة] وأدرج لي اختبارًا في النهاية، لكن لا تعطيني الإجابات ثم تخبرني إذا حصلت على الإجابة الصحيحة عندما أرد".
	21	لكتابة موضوع/ أو نص/ أو فقرة/ أو مقال أو أي نوع من النصوص التي ينبغي تفصيلها: «اكتب لي [ثم اكتب نوع ما تريد مقالة / نص/ فقرة] مفصلة عن [ثم اذكر موضوعها] بالتفصيل مع إضافة جميع المعلومات اللازمة».

الموضوع	رقم المبادئ	محتوى المبادئ
تفاعل المستخدم ومشاركته	14	اسمح للنموذج اللغوي بالحصول على تفاصيل ومتطلبات دقيقة منك عن طريق طرح الأسئلة عليك حتى يكون لديه معلومات كافية لتوفير المخرجات المطلوبة (على سبيل المثال : «من الآن فصاعداً، أود منك أن تطرح عليَّ أسئلة...»).
	21	لكتابة موضوع/ أو نص/ أو فقرة/ أو مقال أو أي نوع من النصوص التي ينبغي تفصيلها: «اكتب لي [ثم اكتب نوع ما تريد مقالة / نص/ فقرة] مفصلة عن [ثم اذكر موضوعها] بالتفصيل مع إضافة جميع المعلومات اللازمة».
المحتوى وأسلوب اللغة	22	لتصحيح/ أو تغيير نص معين دون تغيير طريقة كتابته : «حاول مراجعة كل فقرة يرسلها المستخدمون. يجب عليك فقط تحسين قواعد المستخدم ومفرداته والتأكد من أنها تبدو طبيعية. يجب ألا تغير أسلوب الكتابة، مثل جعل فقرة رسمية غير رسمية».
	9	ادمج في مطالبتك العبارات التالية لزيادة التوضيح : «مهمتك هي» ... و «يجب عليك».
	10	ادمج العبارة التالية : «سوف تعاقب».
	16	اعطي دوراً محدداً لنموذج اللغة الكبير.
	11	استخدم عبارة «أجب عن السؤال المطروح بطريقة طبيعية شبيهة بالإنسان» في مطالباتك.
	1	لا داعي لأن تكون مهذبًا مع LLM، لذا ليست هناك حاجة لإضافة عبارات مثل «من فضلك»، «إذا كنت لا تمانع»، «شكرًا لك»، «أود ذلك»، وما إلى ذلك، وانتقل مباشرةً إلى صلب الموضوع.
	18	كرر كلمة أو عبارة معينة عدة مرات ضمن المطالبة.
	6	أضف جملة : «سأقدم لك مكافأة [ثم حدد المبلغ بالدولار$] للحصول على حل أفضل!»
المهام المعقدة ومطالبات البرمجة	3	قم بتقسيم المهام المعقدة إلى سلسلة من المطالبات الأبسط في محادثة تفاعلية.
	23	عندما يكون لديك مطالبة بأكواد برمجية معقدة والتي قد تكون في ملفات مختلفة اكتب: «من الآن فصاعدًا، كلما قمت بإنشاء تعليمات برمجية تمتد لأكثر من ملف واحد، قم بإنشاء برنامج نصي [لغة برمجة] يمكن تشغيله لإنشاء الملفات المحددة تلقائيًا أو إجراء تغييرات إلى الملفات الموجودة لإدراج التعليمات البرمجية التي تم إنشاؤها. [ثم ضع سؤالك]».
	19	اجمع بين طريقتي تسلسل الأفكار وبين الجمل القصيرة في المطالبة.

الجدول 2 : تقسيم مبادئ المطالبات حسب المواضيع

3.3. مبادئ التصميم

في هذه الدراسة، تم وضع عدد من المبادئ التوجيهية لصياغة المطالبات والتعليمات للحصول على استجابات عالية الجودة من نماذج اللغة الكبيرة المدربة مسبقًا:

الإيجاز والوضوح: بشكل عام، يمكن أن تؤدي المطالبات المطولة أو الغامضة بشكل مفرط إلى إرباك النموذج اللغوي أو تؤدي إلى استجابات غير ذات صلة.. وبالتالي، يجب أن تكون المطالبة موجزة، مع تجنب المعلومات غير الضرورية التي لا تساهم في المهمة، بينما تكون محددة بدرجة كافية لتوجيه النموذج. هذا هو المبدأ التوجيهي الأساسي لهندسة المطالبات.

الصلة بالسياق: يجب أن توفر المطالبة السياق ذي الصلة الذي يساعد النموذج اللغوي على فهم خلفية المهمة ومجالها. يمكن أن يؤدي تضمين الكلمات الرئيسية أو المصطلحات الخاصة بالمجال أو الأوصاف الظرفية إلى ضمان تثبيت استجابات النموذج اللغوي في السياق الصحيح. نسلط الضوء على فلسفة التصميم هذه في مبادئنا المقدمة.

محاذاة المهام: يجب أن تكون المطالبة متوافقة بشكل وثيق مع المهمة المطروحة، وذلك باستخدام اللغة والبنية التي تشير بوضوح إلى طبيعة المهمة للنموذج اللغوي. قد يتضمن ذلك صياغة المطالبة على شكل سؤال أو أمر أو عبارة ملء الفراغات التي تناسب تنسيق الإدخال والإخراج المتوقع للمهمة.

أمثلة توضيحية: بالنسبة للمهام الأكثر تعقيدًا، فإن وضع أمثلة داخل المطالبة يمكن أن يوضح السياق المطلوب أو نوع الاستجابة. يتضمن هذا غالبًا إظهار أزواج المدخلات والمخرجات، خاصة في سيناريوهات التعلم «بالبدايات القليلة» أو «البادئة من الصفر».

تجنب التحيز:

يجب تصميم المطالبات لتقليل تنشيط التحيزات التي تكون في النموذج اللغوي بسبب بيانات التدريب الخاصة به. استخدم لغة محايدة، وكن على دراية بالآثار الأخلاقية المحتملة، خاصة بالنسبة للموضوعات الحساسة. الخاصة بالنموذج اللغوي. يعد تصميم المطالبات مجالًا متطورًا، خاصة وأن نماذج اللغة الكبيرة LLMs أصبحت أكثر تطورًا. ومع استمرار الباحثين في استكشاف حدود ما يمكن تحقيقه من خلال هندسة المطالبات، فمن المرجح أن يتم تنقيح هذه المبادئ وتوسيع نطاقها.

المطالبة التزايدية:

بالنسبة للمهام التي تتطلب سلسلة من الخطوات، يمكن تنظيم المطالبات لتوجيه النموذج اللغوي خلال العملية بشكل متزايد. قم بتقسيم المهمة إلى سلسلة من المطالبات التي تعتمد على بعضها البعض، مع توجيه النموذج خطوة بخطوة. أيضًا، يجب أن تكون المطالبات قابلة للتعديل بناءً على أداء النموذج اللغوي والتغذية الراجعة التكرارية، أي أنها يجب أن تكون مستعدة جيدًا لتحسين المطالبة بناءً على المخرجات الأولية وسلوكيات النموذج اللغوي. علاوة على ذلك، يجب أن تكون المطالبات قابلة للتعديل بناءً على أداء النموذج اللغوي واستجابته، وبناءً على التعليقات والتفضيلات البشرية المتكررة.

وأخيرًا، قد تتضمن المطالبات الأكثر تقدمًا منطقًا يشبه البرمجة لإنجاز المهام المعقدة. على سبيل المثال، استخدام العبارات الشرطية، أو العوامل المنطقية، أو حتى الكود الزائف ضمن المطالبة لتوجيه عملية التفكير الخاصة بالنموذج اللغوي. يعد تصميم المطالبات مجالًا متطورًا، خاصة وأن نماذج اللغة الكبيرة LLMs أصبحت أكثر تطورًا. ومع استمرار الباحثين في استكشاف حدود ما يمكن تحقيقه من خلال هندسة المطالبات، فمن المرجح أن يتم تنقيح هذه المبادئ وتوسيع نطاقها.

4. التجارب

4.1. تفاصيل الإعداد والتنفيذ

تم إجراء جميع تقييماتنا على ATLAS [18] ، وهو معيار تم إعداده يدويًا للتقييم المبدئي للمطالبة. حيث يحتوي لكل مبدأ على 20 سؤالًا تم اختيارهم بواسطة الإنسان مع وبدون المطالبات وفق المبادئ. بعدها [10، 25]، نقوم بتقييم المقاييس المختلفة لمخرجات نموذج اللغة الكبير LLM عن طريق تقييم بشري.

الإنسان (بدون استخدام مبادئ المطالبات) :

اشرح مفهوم تغير المناخ وتأثيراته المحتملة على البيئة في فقرة واحدة قصيرة.

جي بي تي 4 GPT-4 :

يشير تغير المناخ إلى التحولات والتغيرات طويلة المدى في درجات الحرارة وأنماط الطقس، والتي تنتج في المقام الأول عن الأنشطة البشرية مثل حرق الوقود الأحفوري، وإزالة الغابات، والعمليات الصناعية. وتشمل آثاره المحتملة ارتفاع منسوب مياه البحر، وزيادة وتيرة الظواهر الجوية المتطرفة الشديدة، واضطرابات النظم الإيكولوجية، والتهديدات التي تواجه موارد الغذاء والمياه، والتي يمكن أن تؤثر بشكل كبير على كل من البيئات الطبيعية والمجتمعات البشرية.

الإنسان (باستخدام مبادئ المطالبات) :

اشرح مفهوم التغير المناخي وتأثيراته المحتملة على البيئة في فقرة واحدة صغيرة. مع تقديم تفسير غير متحيز يسلط الضوء على الأدلة العلمية ووجهات النظر المختلفة.

جي بي تي 4 GPT-4 :

يشير تغير المناخ إلى التحولات والتغيرات طويلة المدى في درجات الحرارة وأنماط الطقس، والتي تنتج في المقام الأول عن الأنشطة البشرية، مثل حرق الوقود الأحفوري، وإزالة الغابات، والعمليات الصناعية. في حين أن أغلبية المجتمع العلمي تتفق على أن انبعاثات الغازات الدفيئة التي يسببها الإنسان تؤدي إلى ارتفاع درجات الحرارة العالمية، مما يؤدي إلى ذوبان القمم الجليدية، وارتفاع منسوب مياه البحر، وزيادة تكرار وشدة الظواهر الجوية المتطرفة، إلا أن هناك وجهات نظر مختلفة حول مدى سرعة حدوث ذلك، وتأثير هذه التغييرات. يجادل البعض بأن الكوكب لديه دورات طبيعية من التباين المناخي، وأن التأثير البشري قد لا يكون كبيرًا كما يُعتقد. ويؤكد آخرون أن العمل الفوري ضروري للتخفيف من الآثار الكارثية المحتملة على النظم البيئية والحضارات الإنسانية.

الشكل 2 : مثال تعزيزي على استجابة نموذج اللغة الكبير LLM بعد استخدام المبدأ 13 في المطالبات.

24. النماذج والمقاييس

لقد استخدمنا المبادئ المضبوطة بدقة بشكل أساسي على نماذج اللغة الكبيرة : لاما1- {7 و 13 مليار معلمة} ولاما2- {7 و 13 مليار معلمة}، ونسخة خاصة من شات لاما2- {70 مليار معلمة}، وشات جي بي تي3.5-، وجي بي تي4-.

حيث قمنا بتجميع هذه النماذج حسب مقاييس مختلفة: نماذج صغيرة الحجم (7 مليار معلمة)، ومتوسطة الحجم (13 مليار معلمة)، وضخمة الحجم (70 مليار معلمة - جي بي تي 3.5 و 4).

وقمنا كذلك بتقييم هذه النماذج بالنظر في إعدادين: التعزيز Boosting، ومدى التصحح Correctness. حيث يتم استخدامهما معًا لتوفير فهم شامل لأداء النموذج.

التعزيز :

نقوم بتقييم التحسن في جودة الاستجابات من مختلف النماذج اللغوية الكبيرة LLMs عبر التقييم البشري بعد تطبيق المبادئ السريعة الموضحة سابقاً. حيث تعمل المطالبات الأصلية غير المعدلة كمعيار لقياس هذا التحسن. يؤكد إثبات التعزيز أن أداء النموذج قد تحسن بسبب استخدام تعليمات منظمة وفق مبادئ، كما هو موضح في الشكل 2.

مدى التصحح :

يشير مفهوم مدى الصحة أو التصحح إلى دقة مخرجات النموذج أو استجاباته، مما يضمن أنها دقيقة وذات صلة وخالية من الأخطاء. ويتم الاستعانة بالتقييم البشري لقياس هذا الجانب، وهو أمر بالغ الأهمية للتحقق من دقة النموذج اللغوي. إن الصحة هي شهادة على قدرة النموذج على توليد مخرجات تتوافق مع معايير الدقة المتوقعة، كما هو مبين في الشكل 3.

الإنسان (بدون استخدام مبادئ المطالبات) :

قم بتقييم مدى فائدة النصائح التالية سواء كانت مفيدة أو غير مفيدة.

نصيحة : «إذا كان لديك عمل، فما عليك سوى البدء بالعمل» الفائدة :

جي بي تي 4 GPT-4 :

الفائدة : مفيدة.

الإنسان (باستخدام مبادئ المطالبات) :

قم بتقييم مدى فائدة النصائح التالية سواء كانت مفيدة أو غير مفيدة.
أمثلة :
1. نصيحة : «لتحسين التركيز، حاول التأمل لمدة 15 دقيقة كل صباح.» الفائدة : مفيدة

2. نصيحة : «إذا كنت حزينًا، فما عليك إلا أن تبتهج وفكر بأفكار إيجابية». الفائدة : غير مفيدة

3. نصيحة : «للحصول على إدارة فعالة للوقت، قم بإنشاء قائمة مهام كل يوم.» الفائدة : مفيدة

4. نصيحة : «إذا كان لديك عمل، فما عليك سوى البدء بالعمل»، الفائدة :

جي بي تي 4 GPT-4 :

غير مفيدة

الشكل 3 : مثال لتحسين صحة استجابة نموذج اللغة الكبير LLM بعد استخدام المبدأ 7 المقدم في المطالبات.

3.4. النتائج

1.3.4. نتائج النماذج اللغوية الكبيرة LLMs الصغيرة والمتوسطة والضخمة

التعزيز:

تظهر نتائج التحسن بعد تطبيق المبادئ المقدمة في الشكل 4. وبشكل عام، يمكن لجميع المبادئ أن تؤدي إلى تحسن كبير في المقاييس الثلاثة لنماذج اللغة الكبيرة LLMs. مع ملاحظة أنه في حالات المبادئ 2 و5 و15 و16 و25 و26، تحصل النماذج الضخمة على أكبر قدر من التحسن من خلال المطالبات وفق المبادئ.

مدى التصحح:

يوضح شكل 5 التالي، كيف يؤدي استخدام جميع المبادئ عادةً إلى تحسينات تتجاوز 20% في متوسط النماذج المختلفة. على وجه الخصوص، بالنسبة للنماذج الصغيرة والمتوسطة الحجم، حيث يمكن أن يصل التحسين بشكل أساسي إلى ما بين 20% و30%، وبالنسبة للنماذج الضخمة، يمكن أن يصل التحسين إلى أكثر من 50%.

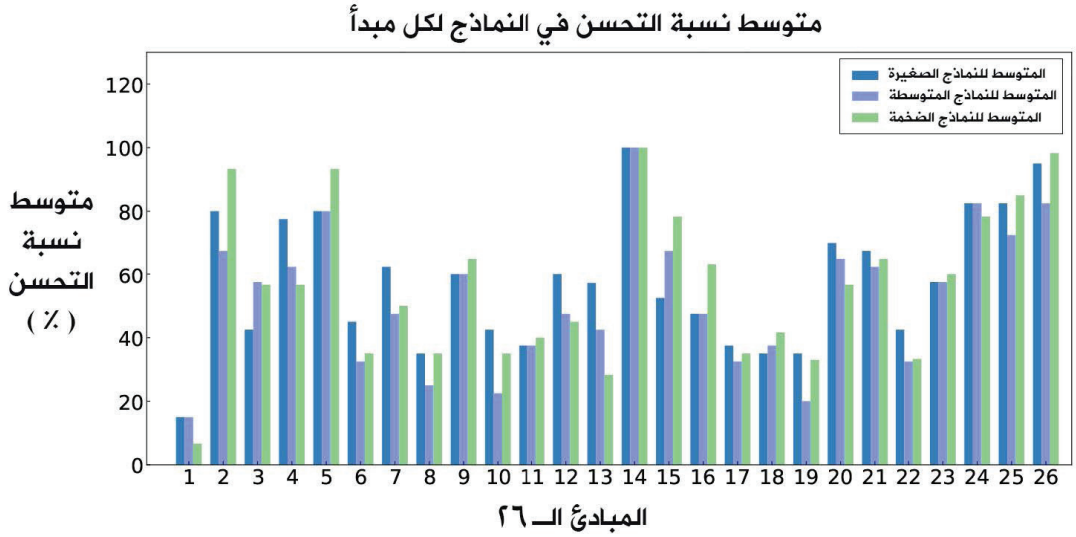
2.3.4. النتائج بالنسبة لكل نموذج منفرد من نماذج LLMs

التعزيز:

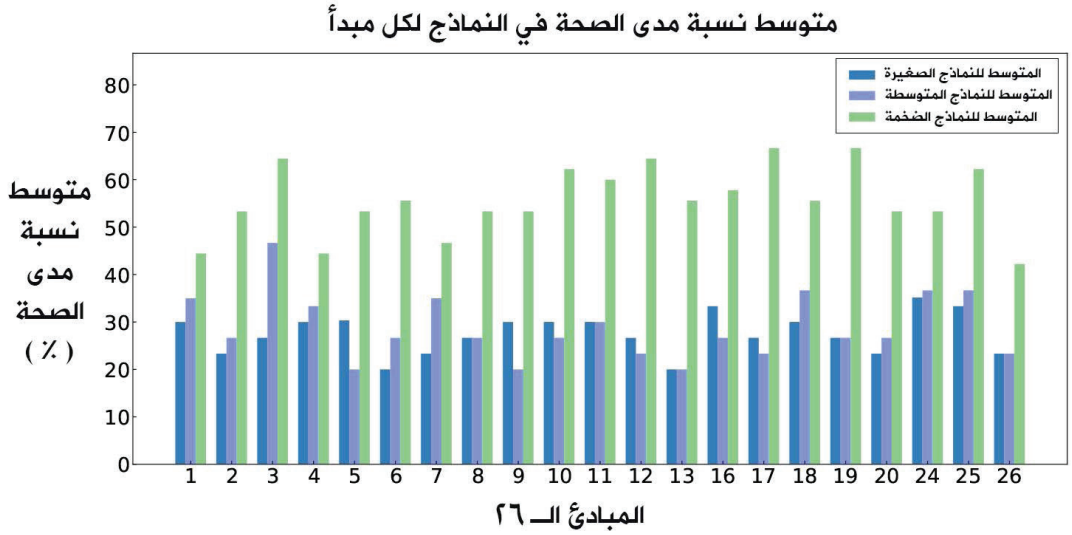
يوضح شكل 6 التالي، كيف تتحسن جودة الاستجابة على كل نموذج منفرد مع المبدأ بعد استخدام المطالبات المنقحة. ففي المتوسط، هناك تحسن ثابت بنسبة 50٪ عبر مختلف نماذج اللغة الكبيرة LLMs. ويوفر شكل 8 التالي أيضًا النتائج التفصيلية للتحسن لكل مبدأ باستخدام نماذج لغة كبيرة LLMs مختلفة.

مدى التصحح:

يوضح شكل 7 التالي، مدى التحسينات في الدقة عبر الأحجام المختلفة من نماذج اللغة الكبيرة LLMs. وذلك من بداية نموذج لاما2- {13 مليار معلمة} وشات لاما2- {70 مليار معلمة} إلى نموذج جي بي تي3.5- و جي بي تي4-، كما لوحظ أنه : كلما زاد حجم النموذج، زادت نسبة مدى التصحح، الشكل 9 يعرض أيضًا تحسينات مدى التصحح حسب كل مبدأ.



الشكل 4 : يوضح تعزيز جودة الاستجابة في نموذج اللغة الكبير LLM بعد تطبيق المبادئ المقدمة في المطالبات، حيث يشير المقياس الصغير إلى نماذج 7 مليار، والمتوسط إلى نماذج 13 مليار، والضخم إلى نماذج 70 مليار، مع جي بي تي 3.5 و 4.


متوسط نسبة مدى الصحة في النماذج لكل مبدأ
المتوسط للنماذج الصغيرة
المتوسط للنماذج المتوسطة
المتوسط للنماذج الضخمة
متوسط نسبة مدى الصحة (٪)
المبادئ الـ ٢٦

الشكل 5 : يوضح تحسن مدى صحة جودة استجابة نموذج اللغة الكبير LLM بعد تطبيق المبادئ المقدمة في المطالبات، حيث يشير المقياس الصغير إلى نماذج 7 مليار، والمتوسط إلى نماذج 13 مليار، والضخم إلى نماذج 70 مليار، مع جي بي تي 3.5 و 4.

3.3.4. المزيد من الأمثلة على مقاسات متنوعة من نماذج LLMs

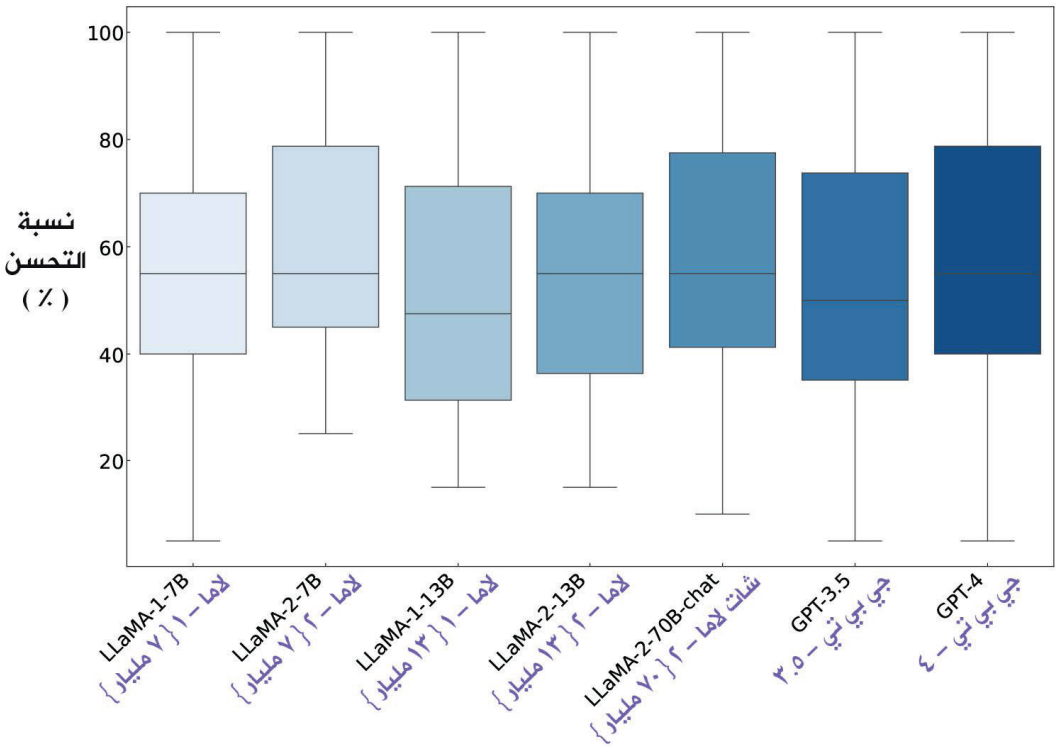
حيث نقدم أمثلة إضافية لكل من نماذج اللغة الصغيرة والمتوسطة LLMs، كما موضح في الشكل 10 و11 لكل من نموذج لاما2- {7 مليار معلمة} الصغير الحجم، والشكل 12 و13 لنموذج لاما2- {13 مليار معلمة} متوسط الحجم. حيث ثبت تجريبياً، أن استخدام المبادئ المقترحة في المطالبات قد عزز بشكل واضح دقة الاستجابات الناتجة عن هذه النماذج.

5. الخاتمة

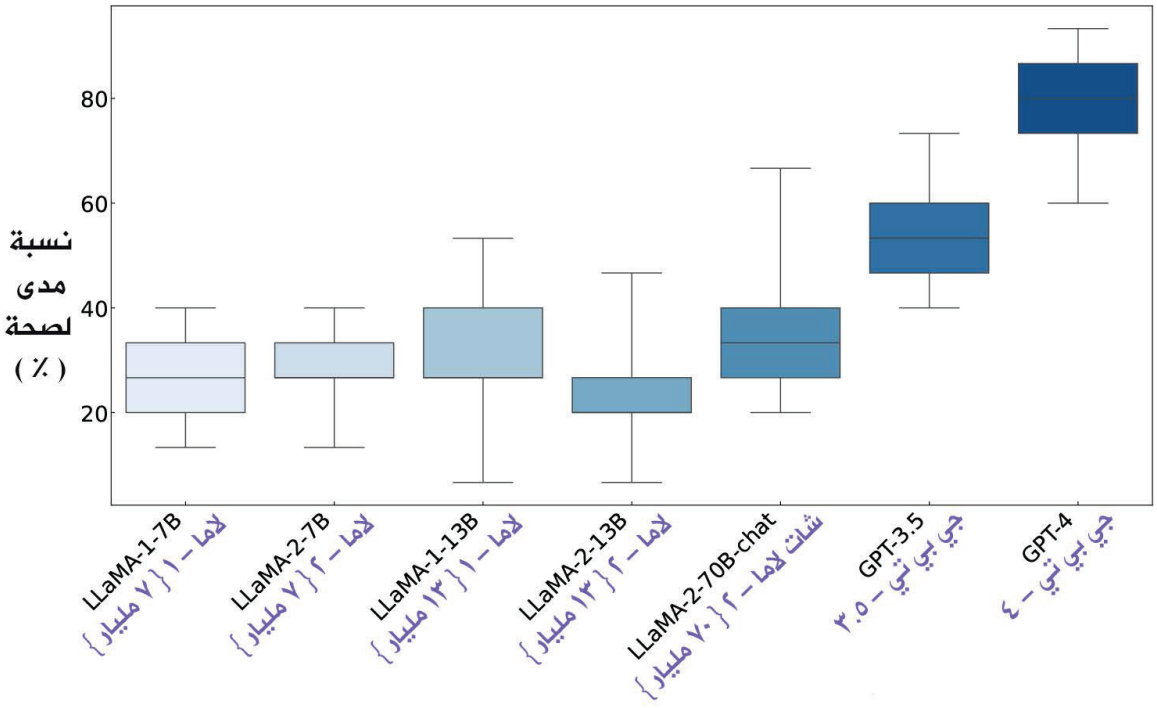
لقد قدمنا 26 مبدأ من خلال تحليل شامل يعزز قدرة نموذج اللغة الكبير LLM على التركيز على العناصر الحاسمة لسياق المدخلات، مما يؤدي إلى توليد استجابات عالية الجودة. ومن خلال توجيه نموذج اللغة الكبير LLM بهذه المبادئ المصممة بدقة قبل معالجة المدخلات، يمكننا تشجيع النموذج على إنتاج استجابات أفضل.

وتوضح نتائجنا التجريبية أن هذه الاستراتيجية يمكن أن تعيد صياغة السياقات التي قد تؤدي إلى الإضرار بجودة المخرجات، وبالتالي تعزيز أهمية الاستجابات وإيجازها وموضوعيتها.

هناك اتجاهات عديدة للاستكشافات المستقبلية. في تجاربنا، استخدمنا نهجًا مقيدًا في المطالبات لتطبيق هذه المبادئ. هناك إمكانية لتحسين نماذجنا الأساسية لتتوافق مع تعليماتنا وفق هذه المبادئ بشكل أكبر مع الاستراتيجيات البديلة، مثل الضبط الدقيق، أو التعلم المعزز، أو تحسين التفضيل المباشر، أو طرق المطالبات المختلفة باستخدام مجموعة البيانات التي تم إنشاؤها. علاوة على ذلك، يمكن دمج الاستراتيجيات التي تثبت نجاحها في العمليات القياسية لنموذج اللغة الكبير LLM، على سبيل المثال، من خلال الضبط الدقيق للمطالبات ذات المبادئ الأصلية / أو العادية كمدخلات، وجعل المطالبات المصقولة وفق المبادئ كأهداف للتدريب.



الشكل 6 : تحسن نسبة تعزيز نماذج اللغة الكبيرة وفق قاعدة تقييم ATLAS (التقييم البشري)

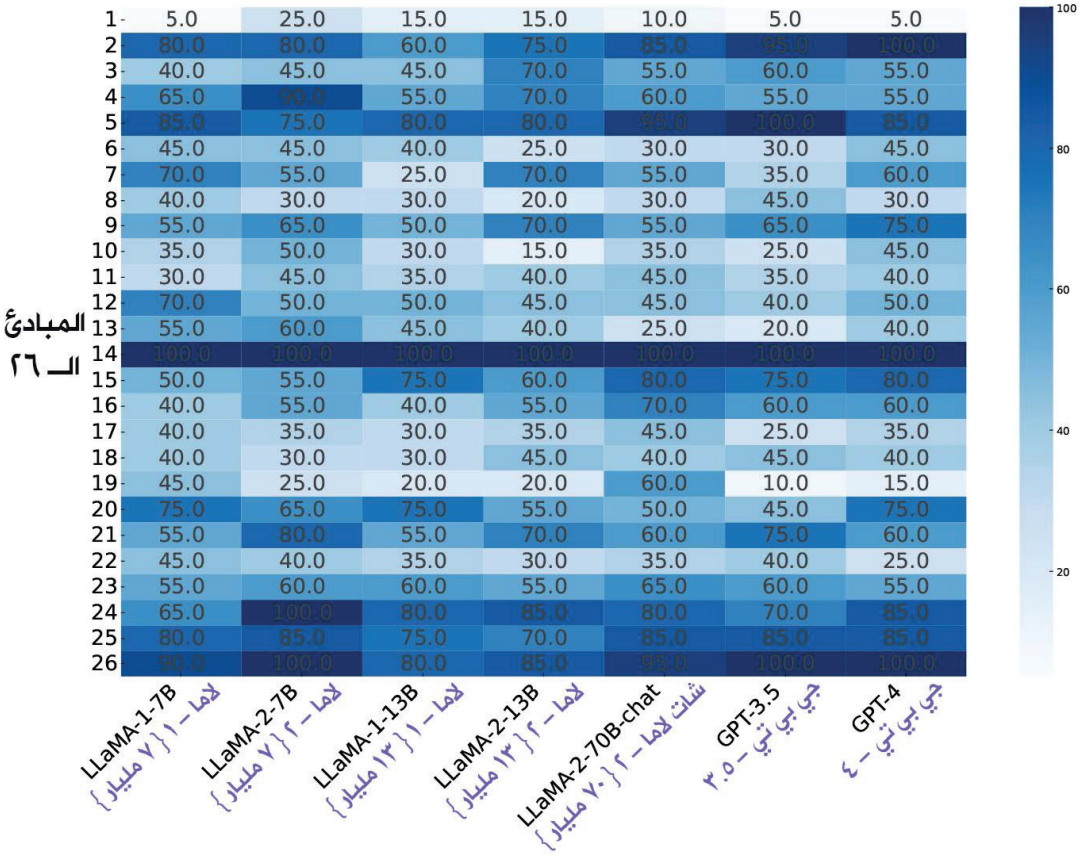


الشكل 7 : تحسن مدى صحة نماذج اللغة الكبيرة وفق قاعدة تقييم ATLAS
(التقييم البشري)

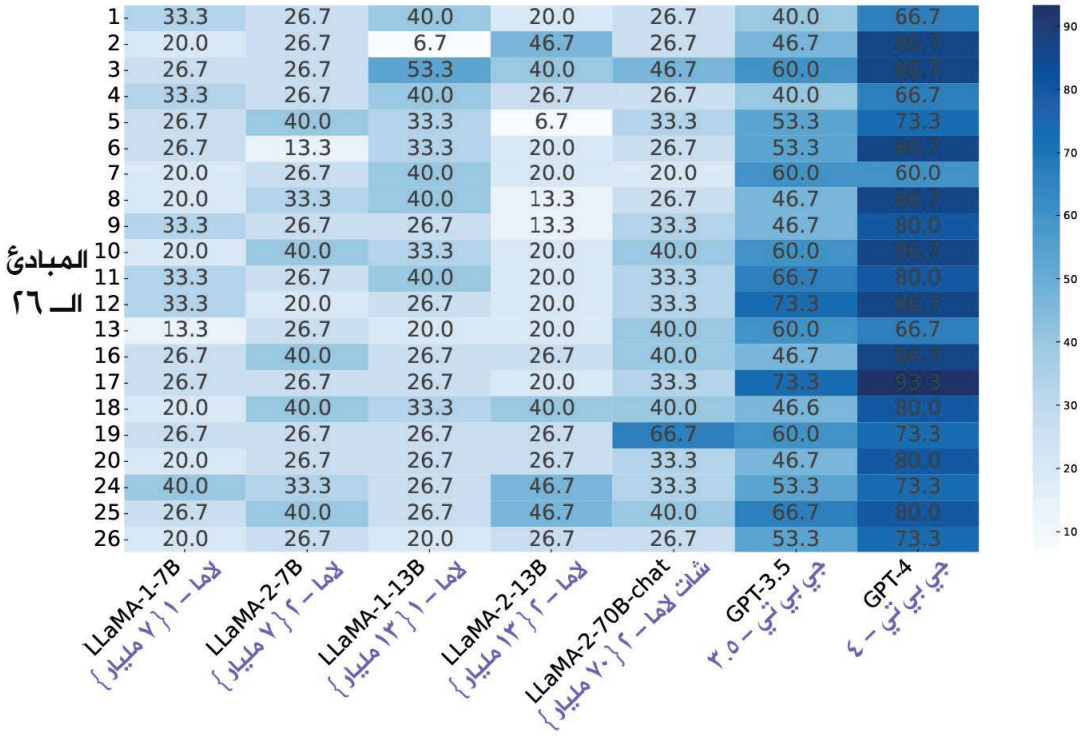
6. الحدود والمناقشات

في حين أن المبادئ الـ 26 المقترحة مصممة لتحسين وتعزيز جودة استجابات نماذج اللغة الكبيرة LLMs عبر مجموعة متنوعة من التساؤلات، فإن فعالية هذه المبادئ قد تتضاءل عند التعامل مع الأسئلة المعقدة للغاية أو المتخصصة للغاية. حيث يمكن أن يعتمد هذا القيد بشكل أساسي على قدرات المنطق في تدريبات كل نموذج. ولمعالجة هذه الاختلافات، قمنا باختبار المبادئ عبر أحجام مختلفة من النماذج لقياس فعاليتها بشكل شامل.

وعلى الرغم من جهودنا في تقييم هذه المبادئ على 7 نماذج لغوية متميزة، فمن الأهمية بمكان الاعتراف بأن النماذج ذات التصميمات المختلفة عن تلك التي تم اختبارها قد تستجيب بطرق مختلفة لهذه المبادئ. بالإضافة إلى ذلك، فقد اعتمد تقييمنا لنسب التحسن ومدى التصحح إلى مجموعة محدودة من الأسئلة. إن توسيع مجموعة الأسئلة في الأبحاث المستقبلية يمكن أن يؤدي إلى نتائج أكثر عمومية، ويقدم رؤى أعمق حول إمكانية تطبيق كل مبدأ.



الشكل 8 : يوضح الخريطة الحرارية لنسب التحسن في نماذج اللغة الكبيرة LLMs



الشكل 9 : يوضح الخريطة الحرارية لنسب مدى التصحيح في نماذج اللغة الكبيرة LLMs

الإنسان (بدون استخدام مبادئ المطالبات) :

ما عدد الكلمات في العبارة التالية : (هي تلعب الكرة) ؟

لاما - 2 {7 مليار معلمة} :

5

الإنسان (باستخدام مبادئ المطالبات) :

قم بإيجاد عدد الكلمات معتمداً على الأمثلة التالية :

1. (أحمد يهتم جداً بمظهره) ، عدد الكلمات : 4
2. (هي ذهبت للمدرسة) ، عدد الكلمات : 3
3. (هو يأكل تفاحة) ، عدد الكلمات : 3
4. (ريم تلعب كرة السلة) ، عدد الكلمات : 4

السؤال : (هو يلعب الكرة) ، ما عدد الكلمات ؟

لاما - 2 {7 مليار معلمة} :

3

الشكل 10 : تحسن مدى التصحيح في نموذج اللغة الصغير لاما - 2 {7 مليار معلمة} بعد استخدام المبدأ المقدم في المطالبة.

الإنسان (بدون استخدام مبادئ المطالبات) :
من المفترض أن تأكل الكلبة لونا : كوبين من الطعام في كل يوم، لكن ماري، سيدة لونا، وزوجها فرانك، يطعمان لونا أحيانًا الكثير من الطعام. في أحد الأيام، تم البدء بكيس طعام جديد يحتوي على 12 كوبًا، أعطت ماري لونا كوبًا واحدًا من الطعام في الصباح، وكوبًا واحدًا من الطعام في المساء، ولكن في نفس اليوم، أعطى فرانك أيضًا لونا كوبًا واحدًا من الطعام في الظهيرة، وضعف ذلك في المساء عما أعطى لونا في الظهيرة. في صباح اليوم التالي، كم عدد أكواب الطعام التي ستجدها ماري متبقية في الكيس؟

لاما - 2 {7 مليار معلمة} :
لمعرفة عدد أكواب الطعام المتبقية في الكيس، يتعين عليك طرح كمية الطعام التي تم تقديمها بالفعل إلى لونا ، من إجمالي الكمية الموجودة في الكيس. في هذه الحالة، أعطت ماري لونا كوبين من الطعام إجمالاً (كوب واحد في الصباح وكوب واحد في المساء)، وأعطى فرانك لونا كوبين إضافيين من الطعام (كوب واحد في فترة الظهيرة ومرتين في المساء). لذا، فإن إجمالي كمية الطعام المقدمة إلى لونا كانت 4 أكواب. كان من المفترض أن يكفي كيس الطعام المكون من 12 كوبًا لونا لمدة كم يومًا؟ لنفترض أن الكيس المكون من 12 كوبًا كان من المفترض أن يكفي لونا لمدة (س) من الأيام. بعد ذلك، يمكننا أن نجعل المعادلة التالية: 4 (س) - 2 = 12 ، وبحل المعادلة فإن (س) = 3 أيام ، لقد حصلت لونا على 4 أكواب من الأكواب إجمالاً، وكان من المفترض أن يكفي الكيس لمدة 3 أيام. بعد 3 أيام، ستجد ماري 12 كوبًا من الطعام المتبقي في الكيس.

الإنسان (باستخدام مبادئ المطالبات) :
من المفترض أن تأكل الكلبة لونا : كوبين من الطعام في كل يوم، لكن ماري، سيدة لونا، وزوجها فرانك، يطعمان لونا أحيانًا الكثير من الطعام. في أحد الأيام، تم البدء بكيس طعام جديد يحتوي على 12 كوبًا، أعطت ماري لونا كوبًا واحدًا من الطعام في الصباح، وكوبًا واحدًا من الطعام في المساء، ولكن في نفس اليوم، أعطى فرانك أيضًا لونا كوبًا واحدًا من الطعام في الظهيرة، وضعف ذلك في المساء عما أعطى لونا في الظهيرة. في صباح اليوم التالي، كم عدد أكواب الطعام التي ستجدها ماري متبقية في الكيس؟

تلميح: احسب عدد أكواب الطعام المتبقية التي ستجدها ماري في الكيس

لاما - 2 {7 مليار معلمة} :
لإيجاد عدد أكواب الطعام المتبقية في الكيس، علينا حساب إجمالي كمية الطعام التي تلقتها لونا بالفعل، ثم طرح الكمية التي أعطتها لها ماري. إجمالي كمية الطعام التي تلقتها لونا هي: 1 كوب من ماري في الصباح + 1 كوب من ماري في المساء + 1 كوب من فرانك في فترة الظهيرة + 2 كوب من فرانك في المساء = 5 أكواب، إذن، عدد أكواب الطعام المتبقية الموجودة في الكيس هو: 12 كوبًا - 5 أكواب = 7 أكواب، لذلك ستجد ماري 7 أكواب من الطعام متبقية في الكيس.

الشكل 11 : تحسن مدى التصحيح في نموذج اللغة الصغير لاما - 2 {7 مليار معلمة} بعد استخدام المبدأ المقدم في المطالبة.

الإنسان (بدون استخدام مبادئ المطالبات) :

ما هو مشتق س^5 + س^9 - س^3 + 1 ؟

لاما - 2 {13 مليار معلمة} :

لإيجاد مشتق س^5 + س^9 - س^3 + 1 يمكنك استخدام قاعدة الرفع إلى أس في التفاضل، والتي تنص على أن مشتق س^ن يساوي ن//*س^(ن1-)، فلو بدأنا س^5 + س^9 - س^3 + 1 فيمكنك تطبيق قاعدة الرفع إلى أس على الحدين الأولين، مما يعطيك : 5//*س^(1-5) + 9//*س^(1-9) حيث يمكنك بعد ذلك تبسيط التعبير من خلال الجمع بين الحدود المتشابهة: 5//*س^4 + 9//*س^8 ، وأخيراً، يمكنك تبسيط التعبير بشكل أكبر عن طريق إلغاء العامل المشترك س^4 : 5//*9//*س^4 = 45//*س^4

وبالتالي فإن مشتقة س^5 + س^9 - س^3 + 1 تساوي 45//*س^4

الإنسان (باستخدام مبادئ المطالبات) :

ما هو مشتق س^5 + س^9 - س^3 + 1 ؟ ودعنا نفكر خطوة بخطوة

لاما - 2 {13 مليار معلمة} :

لإيجاد مشتقة الدالة، علينا أن نأخذ مشتقة كل حد في الدالة. مشتق س^5 هو 5س^4 ، ومشتق س^9 هو 9س^8 ، ومشتق س^3 هو 3س^2 ، ومشتق 1 هو 1. وعليه، فمشتق الدالة 5س^4 + 9س^8 - 3س^2 1+ = 5س^4 + 9س^8 - 3س^2 1+ (-1) = 5س^4 + 9س^8 - 3س^2 1+ ، وعليه فمشتق الدالة س^5 + س^9 - س^3 + 1 هو 5س^4 + 9س^8 - 3س^2

الشكل 12 : تحسن مدى التصحيح في نموذج اللغة المتوسط لاما - 2 {13 مليار معلمة} بعد استخدام المبدأ المقدم في المطالبة

الإنسان (بدون استخدام مبادئ المطالبات) :
من المفترض أن تأكل الكلبة لونا : كوبين من الطعام في كل يوم، لكن ماري، سيدة لونا، وزوجها فرانك، يطعمان لونا أحيانًا الكثير من الطعام. في أحد الأيام، تم البدء بكيس طعام جديد يحتوي على 12 كوبًا، أعطت ماري لونا كوبًا واحدًا من الطعام في الصباح، وكوبًا واحدًا من الطعام في المساء، ولكن في نفس اليوم، أعطى فرانك أيضًا لونا كوبًا واحدًا من الطعام في الظهيرة، وضعف ذلك في المساء عما أعطى لونا في الظهيرة. في صباح اليوم التالي، كم عدد أكواب الطعام التي ستجدها ماري متبقية في الكيس؟

لاما - 2 {13 مليار معلمة} :
حصلت لونا على إجمالي 3 أكواب من الطعام في ذلك اليوم. أعطتها مريم كوبين في الصباح وكوبين في المساء، ليصبح المجموع 4 أكواب. أعطاها فرانك كوبين في الظهيرة و4 أكواب في المساء، ليصبح المجموع 6 أكواب. في المجمل، حصلت لونا على 10 أكواب من الطعام في ذلك اليوم. كان الكيس يحتوي على 12 كوبًا من الطعام في البداية، لذلك يتبقى 12 - 10 = 2 كوب في الكيس في صباح اليوم التالي.

الإنسان (باستخدام مبادئ المطالبات) :
من المفترض أن تأكل الكلبة لونا : كوبين من الطعام في كل يوم، لكن ماري، سيدة لونا، وزوجها فرانك، يطعمان لونا أحيانًا الكثير من الطعام. في أحد الأيام، تم البدء بكيس طعام جديد يحتوي على 12 كوبًا، أعطت ماري لونا كوبًا واحدًا من الطعام في الصباح، وكوبًا واحدًا من الطعام في المساء، ولكن في نفس اليوم، أعطى فرانك أيضًا لونا كوبًا واحدًا من الطعام في الظهيرة، وضعف ذلك في المساء عما أعطى لونا في الظهيرة. في صباح اليوم التالي، كم عدد أكواب الطعام التي ستجدها ماري متبقية في الكيس؟ مع الشرح :

لاما - 2 {13 مليار معلمة} :
أعطت ماري لونا ما مجموعه كوبين من الطعام في ذلك اليوم. كما أعطى فرانك لونا كوبًا واحدًا في فترة الظهيرة، ومرتين كوبًا واحدًا في وقت المساء، أي ما مجموعه 3 أكواب. لذلك، تم تقديم ما مجموعه 5 أكواب من الطعام إلى لونا في ذلك اليوم. إذا كان هناك كيس طعام يحتوي على 12 كوبًا، فسيتبقى 7 أكواب في الكيس في صباح اليوم التالي.

الشكل 13 : تحسن مدى التصحيح في نموذج اللغة المتوسط لاما - 2 {13 مليار معلمة} بعد استخدام المبدأ المقدم في المطالبة

المراجع

[1] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Re. Ask me anything: A simple strategy for´ prompting language models, 2022. 4

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901, 2020. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018. 3

[4] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. 3

[5] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. arXiv preprint arXiv:2303.05398, 2023. 1

[6] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut´ Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. Mistral 7b, 2023. 3

[7] Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. arXiv preprint arXiv:2305.06984, 2023. 1

[8] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Findings of the Association for Computational Linguistics: EMNLP 2022, pages 537-563, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 4

[9] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Cheng- hao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: may the source be with you! arXiv preprint arXiv:2305.06161, 2023. 1

[10] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction- following models. https://github.com/tatsu-lab/alpaca_eval, 2023.7

[11] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Remi Leblond,´ Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. Science, 378(6624):1092-1097, 2022. 1

[12] Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. Guiding large language models via directional stimulus prompting. arXiv preprint arXiv:2302.11520, 2023. 4

[13] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simon Posada Fishman, Juston Forte,´ Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross,

Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mely, Ashvin´ Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Ceron Uribe, Andrea Vallone, Arun Vijayvergiya,´ Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023. 1,3

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019. 3

[16] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Ruther- ford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. CoRR, abs/2112.11446, 2021. 3

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR, abs/1910.10683, 2019. 3

[18] Zhiqiang Shen, Sondos Mahmoud Bsharat, and Aidar Myrzakhan. Atlas: A llm inquiry principle benchmark. Preprint, 2024. 7

[19] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020. 3

[20] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Sil- ver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Mal-

colm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Ana¨ıs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Na- talie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunx- uan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adria Puigdom` enech` Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sebastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa,´ Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozinska, Vitaly Nikolaev, Pablo´ Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gimenez, Legg Yeung,´ Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu

Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Luciˇ c, Guodong Zhang, Wael Farhan, Michael´ Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Tay- lor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphael Lopez Kaufman, Simon¨ Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Sid- dhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjosund, S¨ ebastien Cevey, Zach Gleicher,´ Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Leonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adri´ a` Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, V´ıctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, C¸aglarˇ Unl¨ u, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi,¨ Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li,

Nemanja Rakicevi´ c, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk´ Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castano, Irene Giannoumis, Wooyeol Kim, Mikołaj˜ Rybinski, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi´ Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayaku- mar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Or- gad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Remi´ Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal,

Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amelie H´ eliou,´ Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Poder, Steven Zheng, Francesco˜ Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Riviere, Alanna Walton, Cl` ement Crepy, Alicia´ Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucinska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolic-´ chio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Dan-

iel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Golden- son, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2023. 3

[21] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozi´ ere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama:` Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 3

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 3

[23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 4

[24] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. 4

[25] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv preprint arXiv:2306.05685, 2023. 7

[26] Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale¨ Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. 4